Yokochi et al. Journal of Biomedical Semantics (2016) 7:16
DOI 10.1186/s13326-016-0057-1

Journal of Biomedical Semantics

SHORT REPORT Open Access

Publication of nuclear magnetic resonance experimental data with semantic web technology and the application thereof to biomedical research of proteins

Masashi Yokochi[1], Naohiro Kobayashi[1], Eldon L. Ulrich[2], Akira R. Kinjo[1], Takeshi Iwata[1], Yannis E. Ioannidis[3], Miron Livny[4], John L. Markley[2], Haruki Nakamura[1], Chojiro Kojima[1] and Toshimichi Fujiwara[1*]

Abstract

Background: The nuclear magnetic resonance (NMR) spectroscopic data for biological macromolecules archived at the BioMagResBank (BMRB) provide a rich resource of biophysical information at atomic resolution. The NMR data archived in NMR-STAR ASCII format have been implemented in a relational database. However, it is still fairly difficult for users to retrieve data from the NMR-STAR files or the relational database in association with data from other biological databases.

Findings: To enhance the interoperability of the BMRB database, we present a full conversion of BMRB entries to two standard structured data formats, XML and RDF, as common open representations of the NMR-STAR data. Moreover, a SPARQL endpoint has been deployed. The described case study demonstrates that a simple query of the SPARQL endpoints of the BMRB, UniProt, and Online Mendelian Inheritance in Man (OMIM), can be used in NMR and structure-based analysis of proteins combined with information of single nucleotide polymorphisms (SNPs) and their phenotypes.

Conclusions: We have developed BMRB/XML and BMRB/RDF and demonstrate their use in performing a federated SPARQL query linking the BMRB to other databases through standard semantic web technologies. This will facilitate data exchange across diverse information resources.

Keywords: NMR, BMRB, Database, XML, RDF

Findings

Background

The BioMagResBank (BMRB; http://www.bmrb.wisc.edu/) is a worldwide data repository for experimental and derived data gathered from nuclear magnetic resonance (NMR) spectroscopic studies of biological molecules [1]. For more than 15 years, the BMRB has used and developed the NMR-STAR format based on the STAR format specifications [2–4] for data archiving and data exchange (see Fig. 1a). The NMR-STAR Dictionary, acting as ontology for NMR-STAR data, continues to be improved and expanded to keep up with the needs of the biomolecular NMR community. The most important parameter archived in the BMRB is assigned chemical shifts, which can be used directly to determine protein secondary structure and to assist in the determination of their solution structures, to identify interactions of small molecules with target proteins for drug discovery, and to characterize protein-protein interactions.

As of 2015, the BMRB archive contains more than 10,000 entries, and ~800 new entries are being added each year. This growing biological data archive has raised researchers' interest in integrating the diverse biological information resources to form new hypotheses for understanding biological functions and phenomena. The best approach for enhancing interoperability of the

* Correspondence: tfjwr@protein.osaka-u.ac.jp
[1]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan
Full list of author information is available at the end of the article

a

```
##########################################
#  Molecular system (assembly) description  #
##########################################

save_assembly
    _Assembly.Sf_category                          assembly
    _Assembly.Sf_framecode                         assembly
    _Assembly.Entry_ID                             15400
    _Assembly.ID                                   1
    _Assembly.Name                                 'F153(FTR) cTnC'
    _Assembly.BMRB_code                            .
    _Assembly.Number_of_components                 2
    _Assembly.Organic_ligands                      .
    _Assembly.Metal_ions                           .
    _Assembly.Non_standard_bonds                   no
    _Assembly.Ambiguous_conformational_states      .
    _Assembly.Ambiguous_chem_comp_sites            .
    _Assembly.Molecules_in_chemical_exchange       .
    _Assembly.Paramagnetic                         no
    _Assembly.Thiol_state                          .
    _Assembly.Molecular_mass                       18500
    _Assembly.Enzyme_commission_number             .
    _Assembly.Details                              'F153(FTR) cTnC'
    _Assembly.DB_query_date                        .
    _Assembly.DB_query_revised_last_date           .
```

b

```
<?xml version="1.0" encoding="UTF-8"?>
<BMRBx:datablock datablockName="15400"
  xmlns:BMRBx="http://bmrbpub.protein.osaka-u.ac.jp/schema/mmcif_nmr-star.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://bmrbpub.protein.osaka-u.ac.jp/schema/mmcif_nmr-star.xsd mmcif_nmr-star.xsd">
  <BMRBx:assemblyCategory>
    <BMRBx:assembly entry_id="15400" id="1">
      <BMRBx:details>F153(FTR) cTnC</BMRBx:details>
      <BMRBx:molecular_mass>18500</BMRBx:molecular_mass>
      <BMRBx:name>F153(FTR) cTnC</BMRBx:name>
      <BMRBx:non_standard_bonds>no</BMRBx:non_standard_bonds>
      <BMRBx:number_of_components>2</BMRBx:number_of_components>
      <BMRBx:paramagnetic>no</BMRBx:paramagnetic>
      <BMRBx:sf_category>assembly</BMRBx:sf_category>
      <BMRBx:sf_framecode>assembly</BMRBx:sf_framecode>
    </BMRBx:assembly>
  </BMRBx:assemblyCategory>
```

c

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:owl="http://www.w3.org/2002/07/owl#"
         xmlns:BMRBx="http://bmrbpub.protein.osaka-u.ac.jp/schema/mmcif_nmr-star.xsd"
         xmlns:BMRBo="http://bmrbpub.protein.osaka-u.ac.jp/schema/mmcif_nmr-star.owl#">
   <BMRBo:datablock rdf:about="http://bmrbpub.protein.osaka-u.ac.jp/rdf/bmr15400"
                    rdfs:label="info:bmrb/15400">
      <rdfs:seeAlso rdf:resource="http://bmrbpub.protein.osaka-u.ac.jp/xml/bmr/bmr15400-noatom.xml"/>
      <rdfs:seeAlso rdf:resource="http://www.bmrb.wisc.edu/ftp/pub/bmrb/entry_lists/nmr-star3.1/bmr15400.str"/>
      <rdfs:seeAlso rdf:resource="http://bmrb.pdbj.org/ftp/pub/bmrb/entry_lists/nmr-star3.1/bmr15400.str"/>
      <rdfs:seeAlso rdf:resource="http://bmrb.cerm.unifi.it/ftp/pub/bmrb/entry_lists/nmr-star3.1/bmr15400.str"/>
      <BMRBo:datablockName>15400</BMRBo:datablockName>
      <BMRBo:has_assemblyCategory>
         <BMRBo:assemblyCategory rdf:about="http://bmrbpub.protein.osaka-u.ac.jp/rdf/bmr15400/assemblyCategory">
            <BMRBo:has_assembly>
               <BMRBo:assembly rdf:about="http://bmrbpub.protein.osaka-u.ac.jp/rdf/bmr15400/assembly/15400,1">
                  <BMRBo:of_datablock rdf:resource="http://bmrbpub.protein.osaka-u.ac.jp/rdf/bmr15400"/>
                  <BMRBo:reference_to_entry>
                     <rdf:Description rdf:about="http://bmrbpub.protein.osaka-u.ac.jp/rdf/bmr15400/entry/15400">
                        <BMRBo:referenced_by_assembly rdf:resource="http://bmrbpub.protein.osaka-u.ac.jp/rdf/bmr15400/assembly/15400,1"/>
                     </rdf:Description>
                  </BMRBo:reference_to_entry>
                  <BMRBo:assembly.entry_id>15400</BMRBo:assembly.entry_id>
                  <BMRBo:assembly.id>1</BMRBo:assembly.id>
                  <BMRBo:assembly.details>F153(FTR) cTnC</BMRBo:assembly.details>
                  <BMRBo:assembly.molecular_mass>18500</BMRBo:assembly.molecular_mass>
                  <BMRBo:assembly.name>F153(FTR) cTnC</BMRBo:assembly.name>
                  <BMRBo:assembly.non_standard_bonds>no</BMRBo:assembly.non_standard_bonds>
                  <BMRBo:assembly.number_of_components>2</BMRBo:assembly.number_of_components>
                  <BMRBo:assembly.paramagnetic>no</BMRBo:assembly.paramagnetic>
                  <BMRBo:assembly.sf_category>assembly</BMRBo:assembly.sf_category>
                  <BMRBo:assembly.sf_framecode>assembly</BMRBo:assembly.sf_framecode>
               </BMRBo:assembly>
            </BMRBo:has_assembly>
         </BMRBo:assemblyCategory>
      </BMRBo:has_assemblyCategory>
```

d

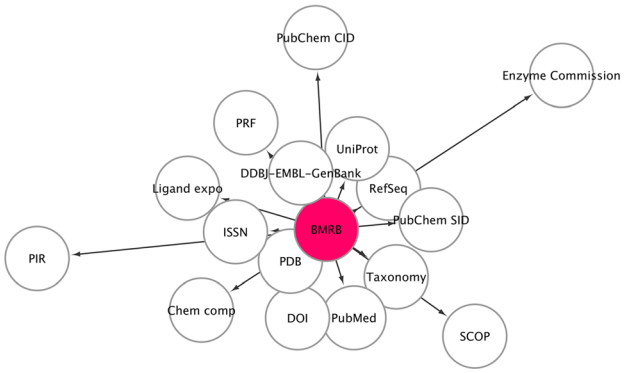

Fig. 1 The sequential data conversion of a BMRB entry from NMR-STAR format. **a** Part of a BMRB entry in NMR-STAR format. Parts of the entry have been converted to **b** XML format and **c** RDF format. **d** Schematic representation of linked external information resources, where shorter distances from the BMRB represent closer relationships with the BMRB

a

```
PREFIX BMRBo: <http://bmrbpub.protein.osaka-u.ac.jp/schema/mmcif_nmr-star.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX omim_v: <http://bio2rdf.org/omim_vocabulary:>

SELECT DISTINCT ?entity_id ?uniprot_id ?label ?omim_id ?dbsnp_id ?mutation ?phenotype
FROM <http://bmrbpub.protein.osaka-u.ac.jp/rdf/bmr>
FROM <http://purl.uniprot.org/uniprot>
FROM <http://bio2rdf.org/omim>
WHERE {
  ?s_up BMRBo:entity_db_link.entry_id "4280" ;
        BMRBo:entity_db_link.entity_id ?entity_id ;
        BMRBo:entity_db_link.database_code "SP" ;
        BMRBo:entity_db_link.accession_code ?uniprot_id ;
        BMRBo:link_to_uniprot ?s_uniprot .
  ?s_uniprot rdfs:label ?info .
  FILTER (STRSTARTS(?info, "info:uniprot"))
  SERVICE <http://sparql.uniprot.org/sparql>
  {
    ?s_uniprot rdfs:label ?label ;
               rdfs:seeAlso ?o_purl .
  }
  FILTER (STRSTARTS(STR(?o_purl), "http://purl.uniprot.org/mim/"))
  BIND (STRAFTER(STR(?o_purl), "http://purl.uniprot.org/mim/") AS ?omim_id)
  BIND (IRI(CONCAT("http://bio2rdf.org/omim:", ?omim_id)) AS ?s_omim)
  SERVICE <http://omim.bio2rdf.org/sparql>
  {
    ?s_omim omim_v:variant ?s_allele .
    ?s_allele omim_v:dbsnp ?s_dbsnp ;
              omim_v:mutation ?mutation ;
              rdfs:label ?phenotype .
    BIND (STRAFTER(STR(?s_dbsnp), "http://bio2rdf.org/dbsnp:") AS ?dbsnp_id)
  }
}
```

b

BMRB	Mutation	OMIM	dbSNP	2nd	SASA%
4280	L100V	300005	rs28935168	Coil	22.5
4280	R106W	300005	rs28934907	Strand	8.0
4280	R133C	300005	rs28934904	Coil	49.4
4280	E137G	300005	rs61748392	Helix	41.7
4280	A140V	300005	rs28934908	Helix	42.0

c

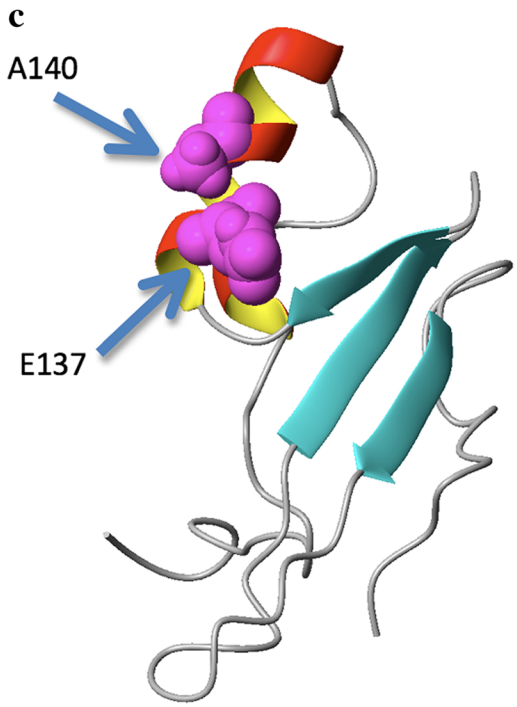

Fig. 2 Federated search using multiple SPARQL endpoints of the BMRB, UniProt, and OMIM. **a** An example of a SPARQL query begins from the BMRB entry. **b** Results performed by the SPARQL query for BMRB entry 4280, showing the BMRB ID, mutation, OMIM ID, dbSNP ID, secondary structure, and SASA. **c** Ribbon models of NMR structure (PDB: 1QK9) of MECP2. The side-chains (space-filling model) for the mutation (E137 and A140) exposed to solvent are responsible for X-linked mental retardation [18]

BMRB would be to convert the archive into standard web formats, XML and RDF, using a data structure that corresponds closely to the NMR-STAR ontology described by an XML schema and OWL.

Methods

We have extended the NMR-STAR Dictionary to accommodate the derived data repositories on BMRB, such as LACS validation reports [5], structural annotations using PACSY [6] and Protein Blocks [7], etc., followed by translation of the dictionary to an XML schema [8] (BMRB/XML Schema), using the PDBx/mmCIF Dictionary Suite developed by the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) (http://sw-tools.rcsb.org/). To automate the XML conversion of BMRB entries, we have developed a software suite (the BMRBxTool) that generates XML documents and validates their format and data consistency according to the BMRB/XML schema. EXtensible Stylesheet Language (XSL) transformation [9] was applied to generate BMRB entries in RDF format from the corresponding XML documents. Along with the RDF, we also provide its ontology written in RDF/RDFS/OWL syntax as BMRB/OWL [10, 11], which is a direct translation of the BMRB/XML schema. To bridge different data models between the XML tree and the RDF directed graph, we have introduced abstract OWL classes and RDF properties to the ontology [12]. For the data conversion from XML to RDF in accordance with the principles of Linked Data [13] and recommendations widely accepted by biological database community [14], we have developed a software suite called BMRBoTool. We have tested as many as thirty SPARQL queries to show how NMR experimental data can be retrieved. In a case study to demonstrate a federated SPARQL query, we performed a search for phenotypes annotated with information for SNPs from the human genome by integrating three SPARQL endpoints: the BMRB, UniProt, and OMIM [see Additional file 1 for details].

Results and discussion

On our portal site (http://bmrbpub.protein.osaka-u.ac.jp/, hereafter abbreviated as '~/'), we have archived the BMRB/XML (~/archive/xml/), as shown in Fig. 1b. The BMRB/RDF derived from the reduced version of the

BMRB/XML (lacking bulky information such as atomic coordinates and NMR restraints) has also been archived (~/archive/rdf/) as shown in Fig. 1c. A bulk download service is available using the *rsync* protocol that helps users mirror the latest data collections, which are updated periodically. A schematic RDF graph of linked databases is illustrated in Fig. 1d. Owing to the machine readability of the XML format, the BMRB/XML provides users with an excellent starting point to develop new tools for use in biochemistry and structural biology. The BMRB/RDF and associated web services enable the integration of the BMRB archive with other biological databases, which may facilitate flexible data exchange and knowledge discovery.

Furthermore, we provide a SPARQL endpoint for querying the BMRB/RDF (~/search/rdf) in the same way as Bio2RDF [15] does. The RDF query language, SPARQL [16], realizes data exchange between databases in a concise syntax. The key feature of SPARQL is its capability of joining remote SPARQL endpoints in what is called a federated SPARQL query [17]. We confirmed the feasibility of such a query (see Fig. 2a), by demonstrating search and classification of SNPs in an associated BMRB entry. In the case study, we successfully collected residues in a BMRB entry whose sequences correspond to SNPs annotated by OMIM, as shown in Fig. 2b. The search results (Fig. 2c) were represented by structural parameters archived in the BMRB [see also Additional file 1 for methods applied and all results, chapter 3], allowing users to infer biological relationships between the phenotype annotated SNPs and structural features in human proteins (see Fig. 2c). The results show that the BMRB/RDF offers a promising approach for integrating biophysical information derived from biological NMR spectroscopy with other bioinformatics resources of interest in biological and medical science research.

Additional file

Additional file 1: Further detail of the BMRB/XML, BMRB/RDF and web services including SPARQL endpoint. Complete results of the SPARQL query (Fig. 2a) and many other SPARQL query examples are also available. (PDF 6814 kb)

Abbreviations

BMRB: BioMagResBank; MECP2: methyl-CpG-binding protein 2; NMR: nuclear magnetic resonance; OMIM: Online Mendelian Inheritance in Man; OWL: web ontology language; RDF: resource description framework; RDFS: RDF schema; SASA: solvent accessible surface area; SNP: single nucleotide polymorphism; SPARQL: SPARQL Protocol and RDF Query Language; XML: eXtensible Markup Language.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MY performed research, data analysis and wrote the paper with NK. TF was principal investigator for this project, contributed to research design and feedback on the manuscript. All authors read and approved the final version of manuscript.

Acknowledgements

This work was supported by National Bioscience Database Center (NBDC) of Japan Science and Technology (JST); and the United States National Library of Medicine [LM05799] and National Institute of General Medical Sciences [GM109046].

Author details

[1]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan. [2]Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA. [3]Department of Informatics & Telecommunications, University of Athens, Athens, Greece. [4]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA.

Received: 11 July 2015 Accepted: 18 March 2016
Published online: 05 May 2016

References

1. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL. BioMagResBank. Nucleic Acids Res. 2008;36:D402–8.
2. Hall SR. The STAR File: a new format for electronic data transfer and archiving. J Chem Inf Comput Sci. 1991;31:326–33.
3. Hall SR, Spadaccini N. The STAR File: Detailed Specifications. J Chem Inf Comput Sci. 1994;34:505–8.
4. Spadaccini N, Hall SR. Extensions to the STAR File Syntax. J Chem Inf Model. 2012;52(8):1901–6.
5. Wang L, Eghbalnia HR, Bahrami A, Markley JL. Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J Biomol NMR. 2005;32:13–22.
6. Lee W, Yu W, Kim S, Chang I, Lee W, Markley JL. PACSY, a relational database management system for protein structure and chemical shift analysis. J Biomol NMR. 2012;54:169–79.
7. Joseph AP, Agarwal G, Mahajan S, Gelly JC, Swapna LS, Offmann B, Cadet F, Bornot A, Tyagi M, Valadié H, Schneider B, Etchebest C, Srinivasan N, De Brevern AG. A short survey on protein blocks. Biophys Rev. 2010;2:137–47.
8. XML Schema Definition Language (XSD) 1.1 Part 1: Structures. 2012, http://www.w3.org/TR/xmlschema11-1/. Accessed 27 October 2015.
9. XSL Transformation (XSLT) Version 2.0. 2007, http://www.w3.org/TR/xslt20/. Accessed 27 October 2015.
10. RDF Schema 1.1. 2014, http://www.w3.org/TR/rdf-schema/. Accessed 27 October 2015.
11. OWL 2 Web Ontology Language Document Overview (Second Edition). 2012, http://www.w3.org/TR/owl2-overview/. Accessed 27 October 2015.
12. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, Kengaku Y, Cho H, Standley DM, Nakagawa A, Nakamura H. Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic Acids Res. 2012;40:D453–60.
13. Berners-Lee T. Linked Data, In Design Issues: Architectural and Philosophical Points. 2006, http://www.w3.org/DesignIssues/LinkedData. Accessed 27 October 2015.
14. Juty N, Le Novère N, Laibe C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res. 2012;40:D580–6.
15. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008;41:706–16.
16. SPARQL 1.1 Query Language. 2013, http://www.w3.org/TR/sparql11-query/. Accessed 27 Oct 2015.
17. DuCharme B. Learning SPARQL. Sebastopol: O'Reilly Media, Inc.; 2011; ISBN: 978-1-449-30659-5.
18. Wakefield RI, Smith BO, Nan X, Free A, Soteriou A, Uhrin D, Bird AP, Barlow PN. The solution structure of the domain from MeCP2 that binds to methylated DNA. J Mol Biol. 1999;291:1055–65.